IS 6733: Deep Learning
Homework 2

P1 (50pt): Write a Python code in Colab using NumPy, Panda, Scikit-Learn to complete the following tasks:

1. Import the Auto MPG dataset with pandas.read_csv() using the dataset URL, use the attribute names as explained in the dataset description as the column names (5pt), view the strings '?' as the missing value, and whitespace (i.e., '\s+') as the column delimiter. Print out the shape and first 5 rows of the obtained DataFrame. (5pt)
   a. Dataset source file: http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data
   b. Dataset description: http://archive.ics.uci.edu/ml/datasets/Auto+MPG
2. Delete the "car_name" column using .drop() method as it is irrelevant to the prediction. Print out a concise summary of the new DataFrame using .info() and check if NULL value exists in each column (5pt)
3. Replace the NULL value with the mean value of the column using .fillna(). Print out the concise summary of the new DataFrame and recheck if NULL value exists in each column (5pt)
4. For the 'origin' column with categorical attribute, replace it with the columns with numerical attributes using one-hot encoding (you can use either get_dummies() in Pandas or OneHotEncoder in Scikit-Learn). Print out the first 5 rows of the newly obtained DataFrame. (10pt)
   a. https://stackabuse.com/one-hot-encoding-in-python-with-pandas-and-scikit-learn/
5. Learn a linear regression model (fit_intercept=True) to predict the "mpg" column from the remaining columns in the obtained DataFrame of Step 4.
   a. Separate the "mpg" column from other columns and view it as the label vector and others as the feature matrix (5pt)
   b. Split the data into a training set (80%) and testing set (20%) using train_test_split and print out their shapes (5pt)
   c. Train the model using the training set and print out the coefficients of the model (5 pt)
   d. Use the learned model to predict on the test set and print out the mean squared error of the predictions (5pt)

P2 (50pt): Write a Python code in Colab using NumPy, Panda, Scikit-Learn to complete the following tasks:

1. Import the red wine dataset with pandas.read_csv() using the dataset URL, use the semi-colon as the column delimiter, and print out both the first five rows and a concise summary of the obtained DataFrame. (10 pt)
   a. Dataset source file http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv
   b. Dataset description: http://archive.ics.uci.edu/ml/datasets/wine+quality

2. Suppose we want to predict the quality of wine from other attributes. Divide the data into a label vector and a feature matrix. Then split them into a training set (80%) and testing set (20%) using train_test_split. (5pt)

3. Use the StandardScalar() in Scikit-Learn to preprocess the feature matrices of both training set and testing set. Note that the testing set can only be scaled by the mean and standard deviation values obtained from the training set. (5 pt)

   a. https://scikit-learn.org/stable/modules/preprocessing.html

4. Use the preprocessed training dataset to learn a classification model with RandomForestClassifier (with 300 trees) in Scikit-Learn. Use 5-fold cross-validation to train and cross-validate the model. Print out the accuracies returned by the five folds as well as their average and standard deviation values. (10 pt)

5. Use GridSearchCV() to find and print out the best hyperparameter for the number of trees (try the following values: 100, 300, 500, 800, 1000 for 'n_estimator'), ignoring the search for other hyperparameters. Also use 5-fold cross-validation during GridSearchCV. (15 pt)

6. Find and print out the testing accuracy of the model obtained using the best hyperparameter value on the preprocessed testing dataset (5 pt)

Submission Instruction: Submit a PDF file of your codes and outputs and a Google Colab shared link to your source file (in the 'comments' box) to Blackboard.